31 | 最先进的数据分析工具展望

2021-10-18 郭炜

《数据分析思维课》 课程介绍 >

讲述：郭炜

时长 12:23 大小 11.35M

数据给你一双看透本质的眼睛，这里是《数据分析思维课》，我是郭炜。

在第27 节课里，我给你分享了我常用的数据分析工具图谱，我们今天再进一步来看看，到底这些工具图谱里最新的技术工具长什么样子？

在这节课里我会给你介绍 2021 年企业里最常用的比较先进的工具，等将来出现了更好的工具，我会在后续的加餐里再做更新。

整个大数据分析框架离不开三个基础技术部分：数据存储、数据处理和数据展示。接下来我就根据这三个基础技术，选三个比较新的开源免费的工具与你分享，我们一块来看看最新的技术可以达到什么的样子。

数据存储与分析引擎——ClickHouse

在数据存储的部分，我给你介绍一下现在全球最流行的专门针对数据分析打造的引擎：ClickHouse。

ClickHouse 专门为数据分析师打造。为什么这么讲呢？

第一，它使用的不是非常复杂的 NoSQL 编程语言，而是用的很简单的 SQL 语言，这样数据分析师、产品经理和运营人员都会非常熟悉。

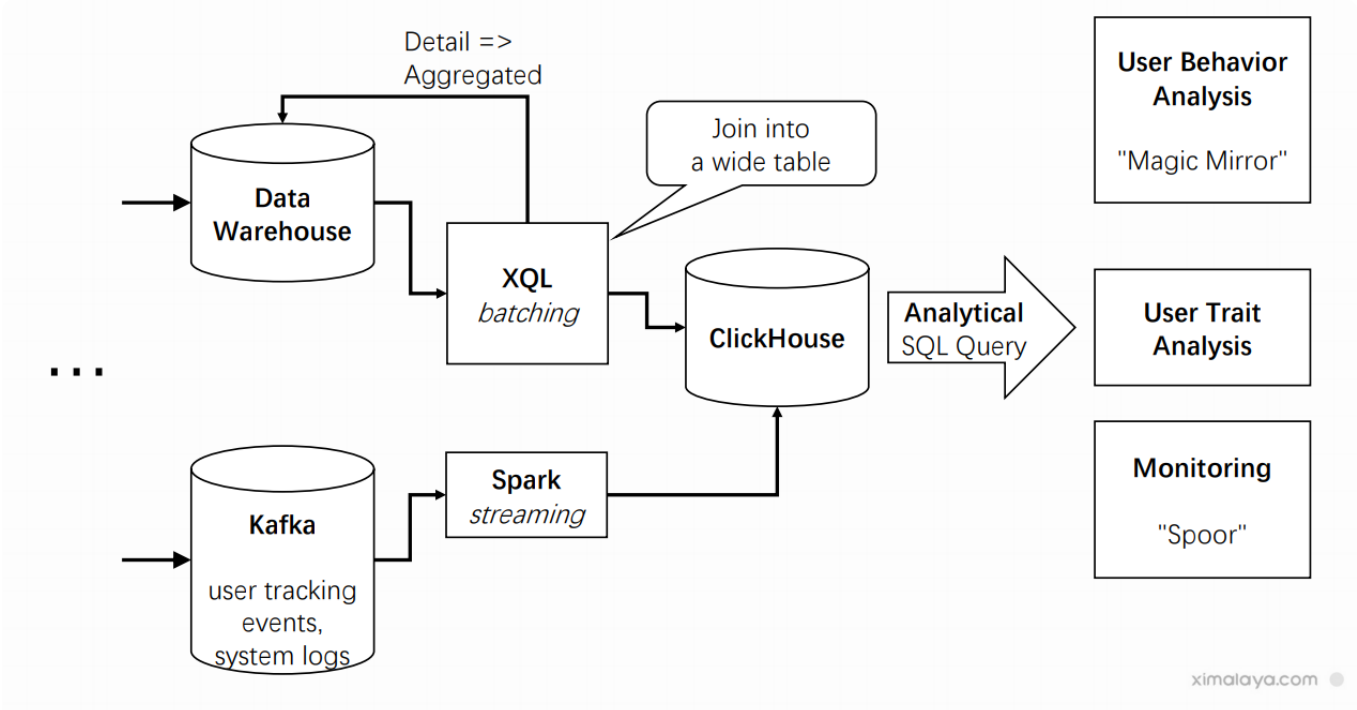
第二，它的宽表查询速度非常快，全球第一。我们做最后数据分析 1 公里的时候，大多数情况都可以用一个或者几个大宽表来解决问题。传统的大数据工具因为要适配各种情况，经常在数据量增大时，整体数据处理的速度变得非常慢。你往往是要填写一个需求单给数据开发和工程部门，让他们转化成复杂的编程语言，或者在大数据平台上面提交一个任务才行。这样一来少则半个小时，多则数天你才可以拿到你想要的结果。

而 ClickHouse 这个引擎数秒之内就可以针对百亿条的数据进行复杂的 SQL 查询。无论是我们需要做分组聚合、明细过滤还是基于文本的条件筛选，它的速度都是秒甚至毫秒级别出结果。这样你在做数据分析的时候，就可以不停地快速进行数据探查，不会打断你的数据分析思路。而不是在提交各种各样的任务后，喝杯咖啡甚至休息几小时之后再看到结果。

第三，ClickHouse 整体部署和维护安装比较简单，在数据量不是特别大的时候，一台服务器就可以搞定，普通的运维人员就可以维护。对于更复杂的情况，你可以使用集群版或者相关的商业版本来提高维护效率。所以 ClickHouse 这个数据引擎特别适合做数据分析，现在已经成为互联网公司的标配数据分析引擎了，下图的这些公司都在使用 ClickHouse 作为它的分析引擎来分析数据。

那 ClickHouse 到底怎么来用的呢？我这里举三个公司对外分享的实例，给你分享一下 ClickHouse 都能干什么。

首先是喜马拉雅的例子，这是 2019 年喜马拉雅 Alex 黄在 ClickHouse Meetup 的时候分享的案例。你能看到在喜马拉雅，ClickHouse 作为一个典型的数据查询平台来使用，它既做用户行为分析，也就是我们网上的各种各样的 APP 和网络日志的查询；也做用户画像的数据分析，你可以对不同用户画像标签的圈选人群进行人群探查、投放效果预测；它还可以用于各种服务器日志的监控报警，当服务出现问题的时候管理可以快速找到相关问题的原因，整体架构图可以参考下面这个样子。



喜马拉雅 Alex，Huang 2019.10.27 ClickHouse Shanghai Meetup

下面一个例子我们来看看腾讯音乐是怎么使用 ClickHouse 的。

腾讯音乐把 ClickHouse 作为了实时分析数据仓库，你在使用腾讯音乐的推荐和点击的时候，背后的大数据平台就是 ClickHouse 来提供的。

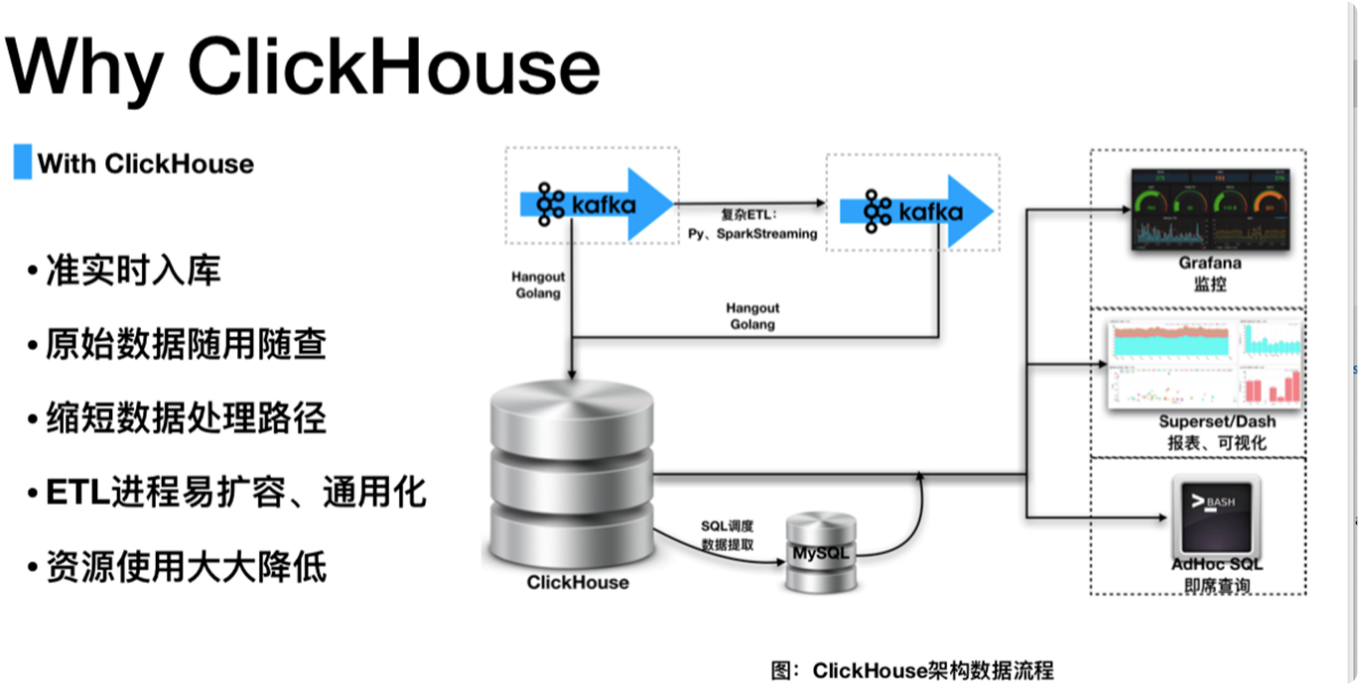
腾讯把数据放到消息队列里，然后通过一个叫做 Flink 的工具实时装载到 ClickHouse 当中，同时把一些离线文件传入传统的数据仓库里。最终我们数据分析师使用的数据，看到的是实时的数据，既可以看到上一秒的系统情况，还可以做各种自定义的 SQL 查询，数据还是秒回。这样帮助腾讯音乐自助进行汇总、筛选查询，也能快速地响应各种各样的原始数据变更。

其实所有互联网大厂里使用 ClickHouse 都可以实现针对数据用户日志的秒级查询，不再需要数据运营和产品团队自己跑复杂的脚本任务和处理了。

腾讯音乐 吴泽君 2021.2.6 On-line Meetup

最后一个例子我们来看看新浪。ClickHouse 在新浪用于监控整个数据平台。这个例子有意思的地方是新浪每天有 300 亿条的数据直接进入到 ClickHouse 平台里，而新浪在做监控的时候是通过算法来进行监控和处理的，每日有 800 万次的查询，每次几乎是毫秒级返回。

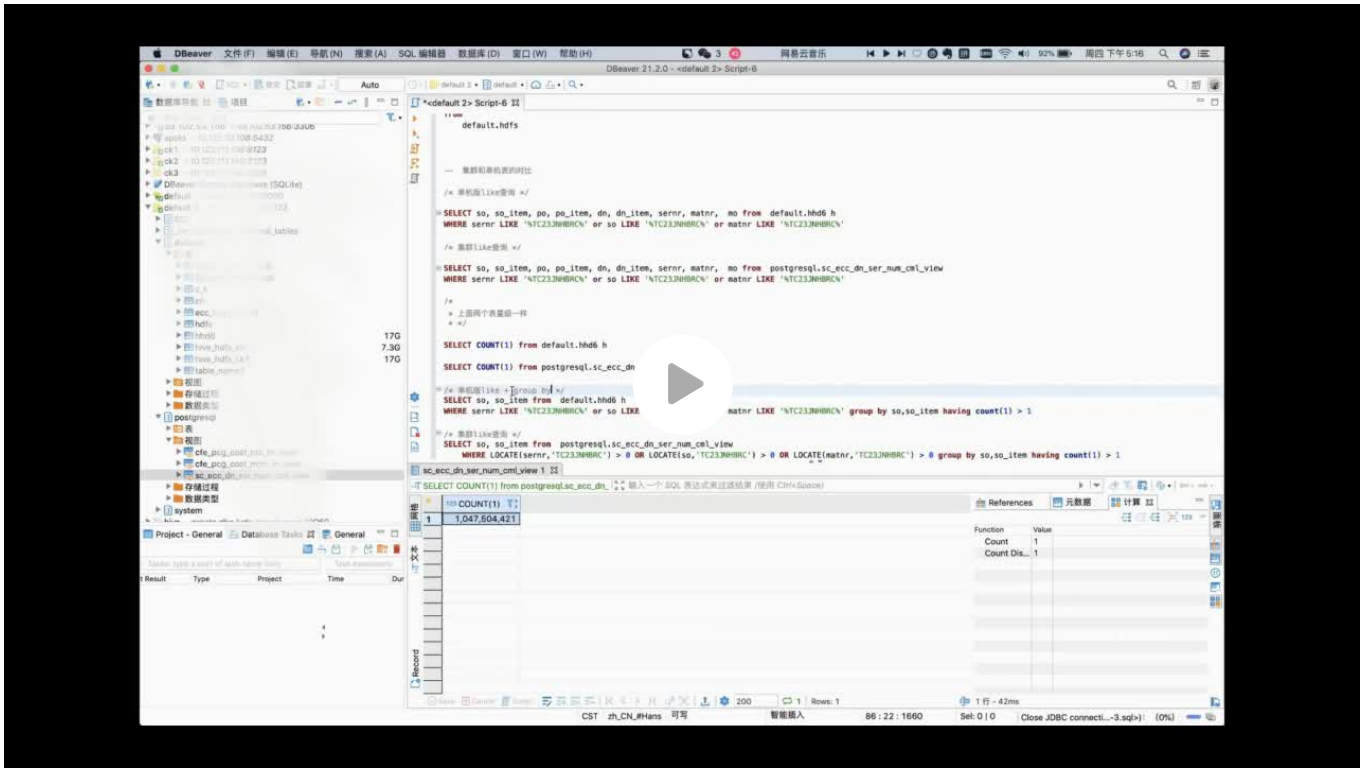
这就像前面数据算法基础这一章提到的，开始时还是用人在做分析，当技术发展到一定程度，我们就可以通过算法和机器来进行分析。现在的数据底层的技术可以已经可以非常容易地做到用算法取代人工，最终高效地实现整体的数据分析和告警。



新浪 JackGao 2019.10.27 ClickHouse Beijing Meetup

数据存储还有很多各种各样的引擎，比如 Spark SQL、Presto、Impala 等等，这些引擎都可以很好地处理数据分析，我只是介绍其中的一种让你体验下最新的数据存储技术是什么样子。

每个数据分析师可以根据自己数据团队的情况选择最合适自己公司的开源产品来使用，我录了一个视频，让你体会下 ClickHouse 到底有多快，一台机器，10 亿条数据复杂查询 4 秒就出结果，而同样的数据 Presto 集群（若干台机器）还要 40 秒才可以出结果，你要是感兴趣可以关注 ClickHouseGroup 微信号来看到最新的 ClickHouse 中国社区 Meetup 信息和中文材料。



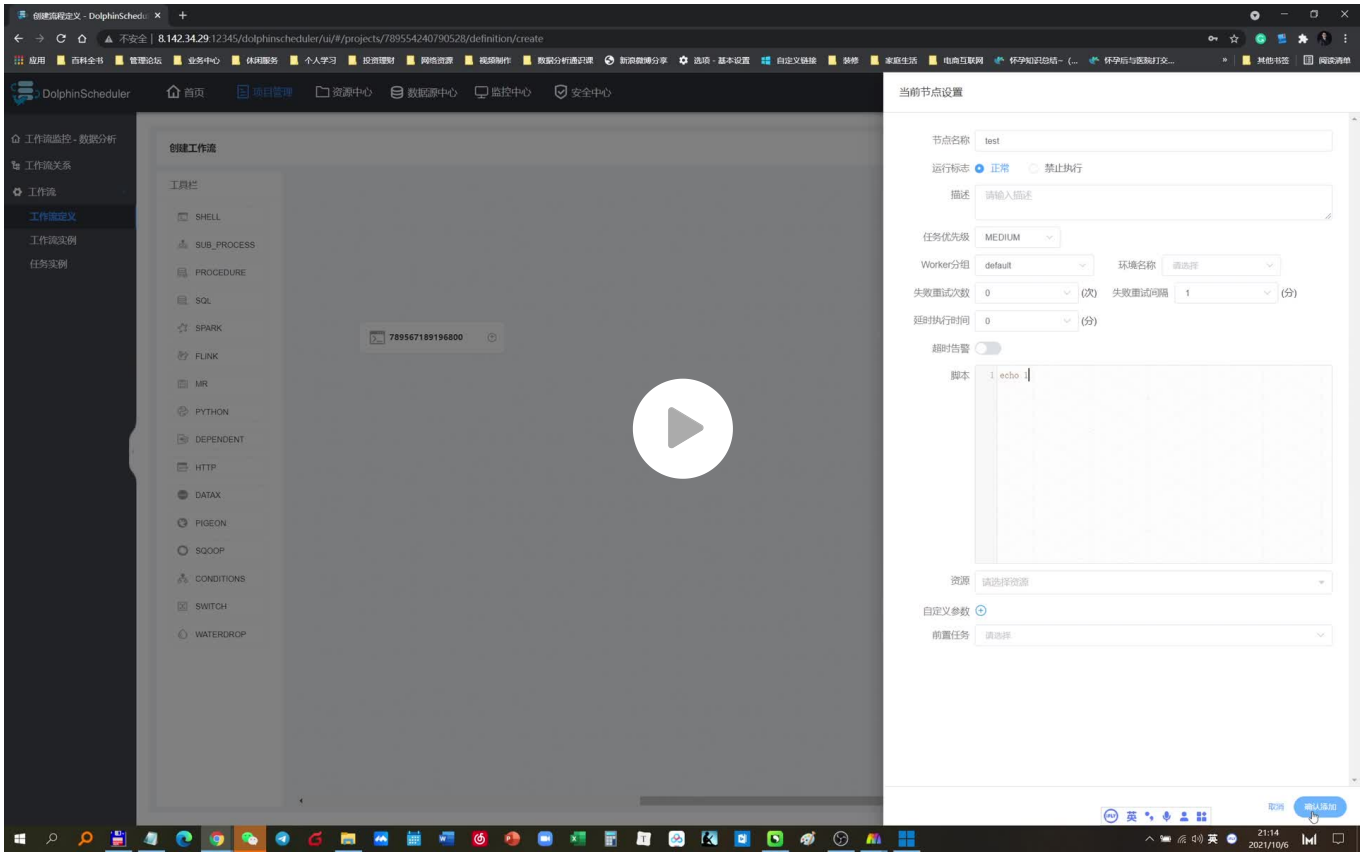
数据处理与调度平台——Apache DolphinScheduler

数据我们都已经存储好了，SQL 你也写好了，如果这是一个日常都要做的任务，你肯定不希望每次都需要人来运行，它要可以自动调度把数据处理到我们目标的数据表里才是最好。这个时候就需要数据运营平台的数据调度和处理引擎了。

这里我给你介绍一个我当年主导开源的数据处理与调度平台，它也是一个 Apache 顶级项目：Apache DolphinScheduler。

它源于我上一家公司易观给内部数据分析师和数据开发人员使用的调度工具。它的优点是全部是可视化的配置（数据分析师最爱），而且超级稳定易扩展。在数据平台人员安装完成之后，你不需要懂得服务器脚本或任何的大数据平台语言，你只需要拖拽把你熟悉的 SQL 脚本放到这个调度平台里，把这些表之间的逻辑关系通过连线把它连起来，你就可以得到一个非常方便使用的大数据调度处理流程了。

底层它采用的是云原生的技术，扩展性和稳定性都是非常优秀的，所以它也是 Apache 基金会的顶级项目。它在我们数据逻辑脚本比较清晰的情况下，可以不用写代码直接配置。我在下面这个视频演示了几个复杂的脚本之间有逻辑关系，怎么样可以不用编程序来快速实现自动化调度启停。



Apache DolphinScheduler 也得到了很多的用户使用。

其中一个互联网用户叫做“奇安信”，他是这样反馈的：Apache DolphinScheduler 是一个可视化非常好的工具，无需代码就可以调度任务。降低使用门槛；它底层用了分布式、易扩展的方式实现了集群高可用；所有的资源文件都是在线的，不用登陆带服务器就可以看到日志错误调试脚本、管理上传的脚本文件；同时它的支持多租户也支持权限管理分给不同部门使用。

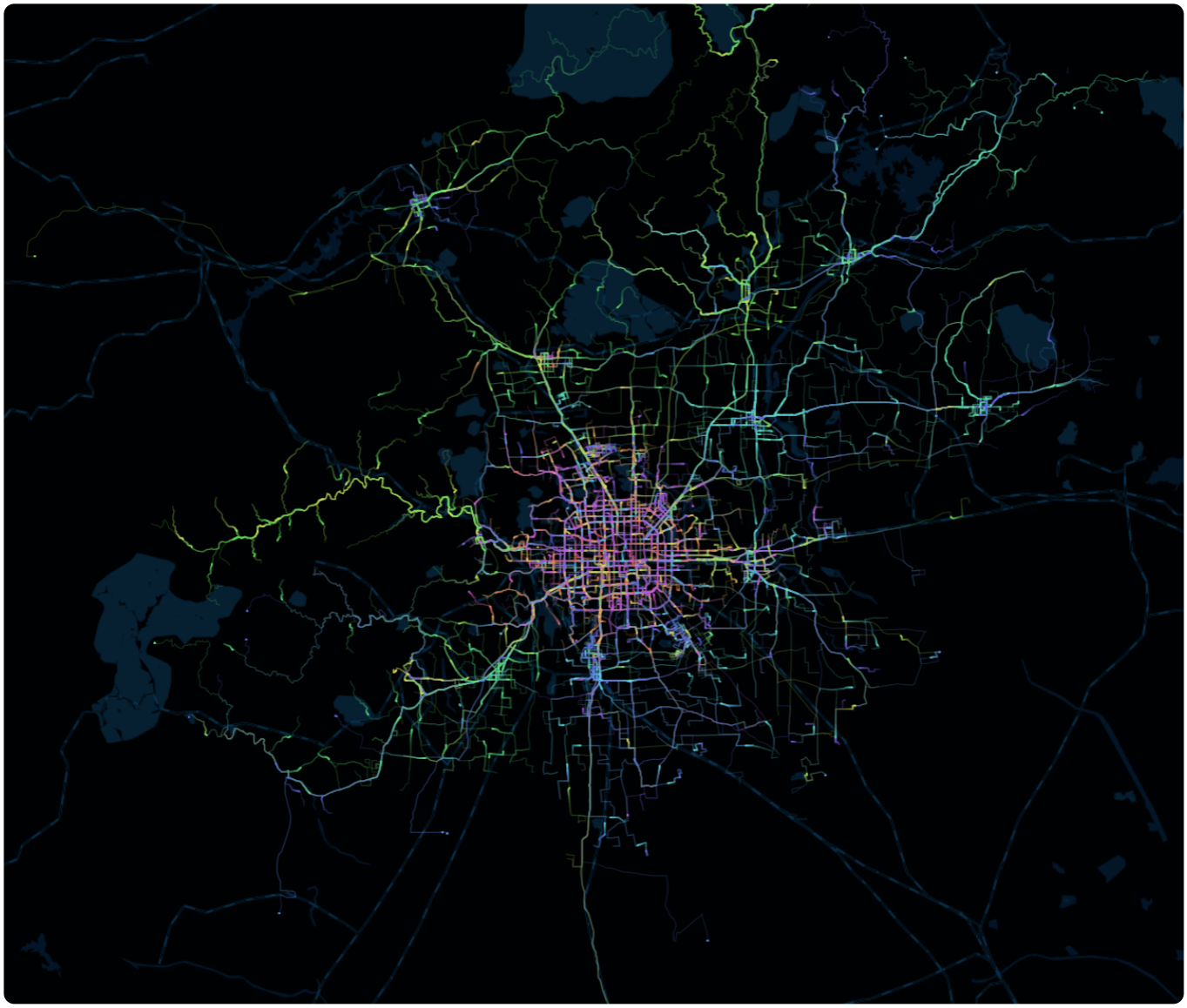
总之 Apache DolphinScheduler 是一个非常好的数据处理与调度工具，你要是感兴趣可以访问官网来看最新的信息。

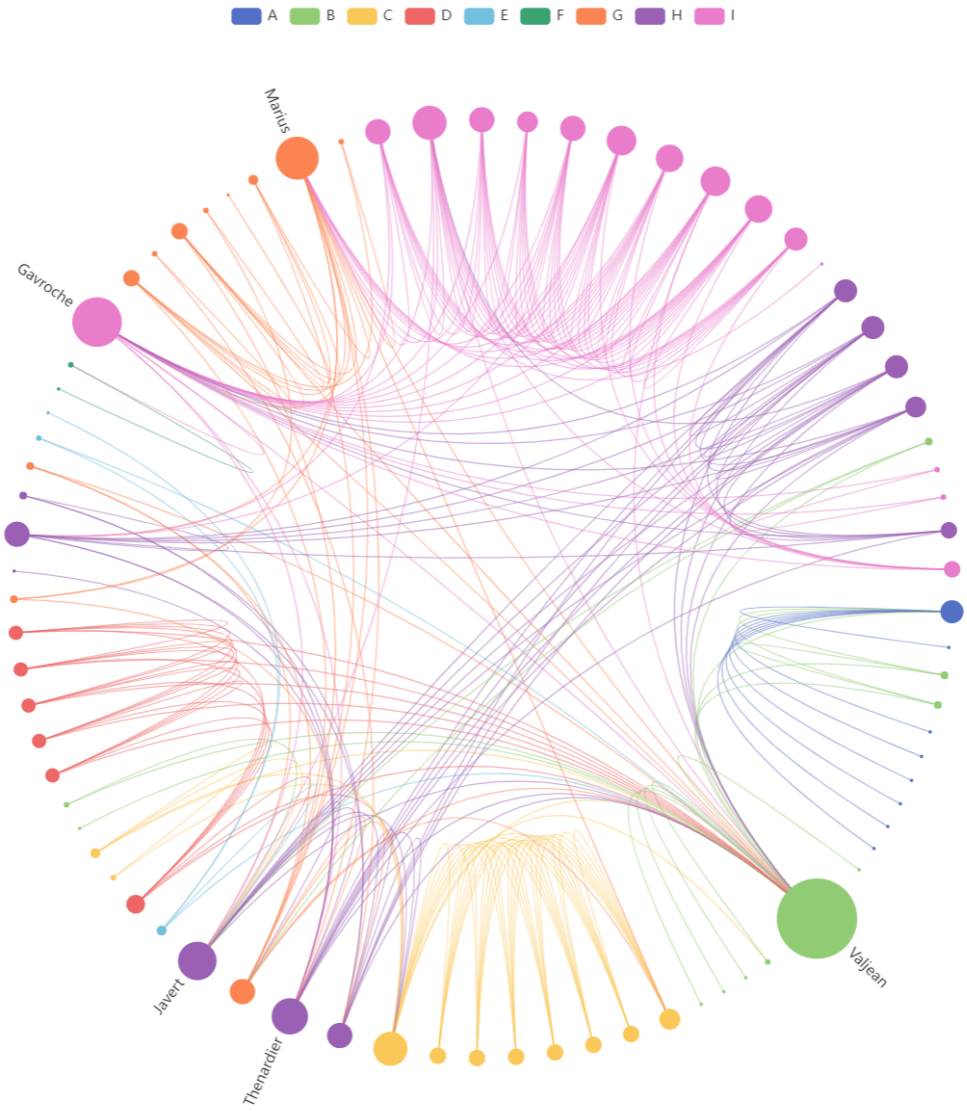
除了 Apache DolphinScheduler，类似的还有美国 AirB&B 开源的 Apache Airflow 和 Cloudera 开源的 Apache OOzie 等等，这些都还需要编一部分代码来执行相关的任务，所以**你可以根据自己的情况去测试一下哪个工具最适合你的场景，最后通过数据平台部门来帮你实现。**

数据展示工具——EChart

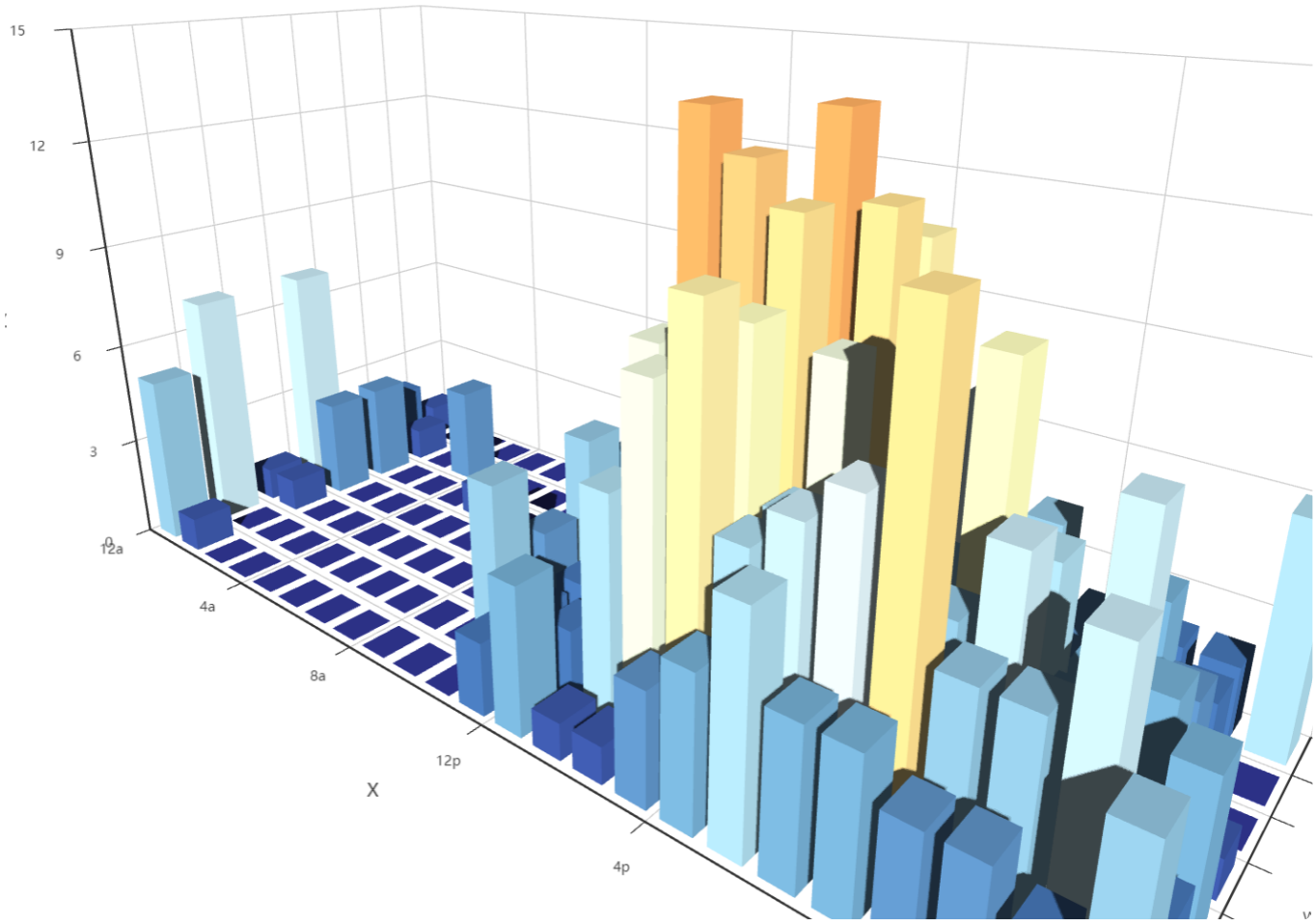
数据存储好了也处理好了，最后还需要非常方便地展示出来，这就不得不提国内的最牛的数据展示工具——Apache ECharts。在前面的课程里我们看过 Excel 各种展示图形的方式，如果你觉得还不够，你可以看一下 ECharts 这个图形展示工具，它是目前市面上最全的展示工具之一。

只不过如果它要变成具体可使用的界面，还需要你请前端小伙伴编一些程序，不过看下图这么酷炫的展示，我觉得你还是值得去要一些前端开发资源的。

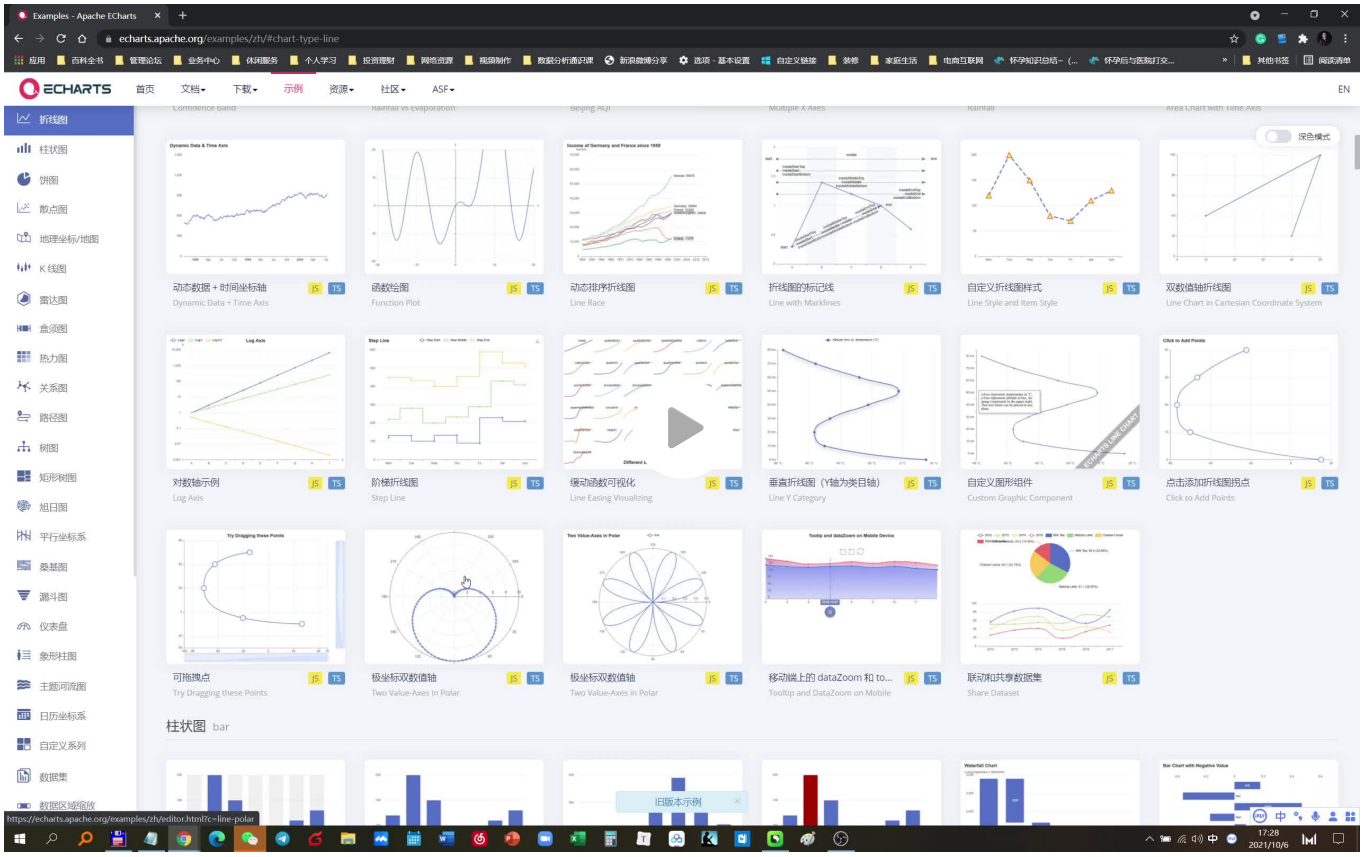




Les Miserables



我在下面这个视频里面给你介绍一下所有 ECharts 支持的图像，你能看到非常全面，几乎你所用数据想要的结果在这里都可以找到。而且它大部分图形支持手机演示，也就是你的数据可以在手机上给所有相关人员展示。同样，你如果感兴趣可以点击这里，访问官网了解更多信息。



当然类似像 Echarts 这样的工具，海外还有 D3 或者是一些其他的收费工具，数据分析最后的颜面还是要找一个漂亮的图形让更多人喜爱。

小结

今天给你介绍了几个比较先进的数据分析工具，你应该能感觉到现在的技术已经非常发达了，只有你想不到，没有做不到的技术工具——无论是单机版实时数据分析引擎，还是无代码的数据逻辑工作流，或是动态酷炫的数据效果，现在都可以通过比较简单的程序化方式实现。

我们现在需要做的是用第三章如何用数据说话的方法捋好自己的数据分析思路和内容，用我们第二章讲到的数据算法基础扩展数据结论，同时你要注意第一章数据分析基础的内容，不要错用数据分析基础知识。最终，你就可以通过这些工具展现一个完整的动态数据分析系统，给公司老板和大多数的用户一起来使用。

基于正确的数据分析思路，使用这些工具打造的数据平台就像一个放大器，可以帮我们快速推广想法，能够帮你将已成型的数据分析思维放大给更多的人迅速使用。

需要注意的是，我们不要沉浸在先进的工具里不能自拔，工具永远只是我们拓展数据实践半径让更多的人认可数据思维的手段，所以你会发现在工具这一部分，我没有给你推荐非

常多酷炫的工具，因为我觉得**对数据分析思维来说，最重要的还是思维。具体用什么武器，其实只要顺手，它哪怕是一个小小的 Excel，也可以使出非常好的数据分析效果来。**如果你没有好的数据分析思维，你的展示再酷炫也是没有意义的。

数据给你一双看透本质的眼睛，数据分析核心在思维，酷炫的工具就像是美颜滤镜，底子好不好，见到真人还是要原形毕露的，所以多花心思在数据分析思路上。

课后思考

你还看到过什么顺手的数据分析工具？分享出来我们一起提高。

分享给需要的人，Ta订阅后你可得 20 元现金奖励

生成海报并分享

赞 1 提建议

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 30 | 快速实现数据分析基础课中的分析模型

下一篇 编辑手记 | 让生活多一些科学，少一些神学

精选留言 (1)

写留言

80分

2021-10-18

这一节介绍了ClickHouse、Apache DolphinScheduler和ECharts。前面两个工具都没接触过，还有很多的东西要学呀。

展开